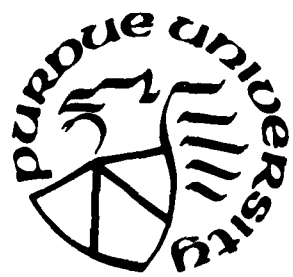
AD-A101 921 PURDUE UNIV LAFAYETTE IN DEPT OF STATISTICS F/G 12/1
SOME SEQUENTIAL SELECTION PROCEDURES FOR GOOD REGRESSION MODELS--ETC(U)
JUN 81 T HSU, D HUANG N00014-75-C-0455
UNCLASSIFIED MMS-81-21 NL

LEVEL II

12

PURDUE UNIVERSITY



DEPARTMENT OF STATISTICS

DIVISION OF MATHEMATICAL SCIENCES

(12)

SOME SEQUENTIAL SELECTION PROCEDURES FOR GOOD REGRESSION MODELS*

by

Tong-An Hsu
National Central University, Chung-Li, Taiwan

and

Deng-Yuan Huang
Academia Sinica, Taipei, Taiwan

Department of Statistics
Division of Mathematical Sciences
Purdue University
Mimeograph Series #81-21

June 1981

*This research was supported by the Office of Naval Research contract N00014-75-C-0455 at Purdue University. Reproduction in whole or in part is permitted for any purpose of the United States Government. The authors would also like to thank the National Science Council of Republic of China for some financial support.

SOME SEQUENTIAL SELECTION PROCEDURES FOR GOOD REGRESSION MODELS

Tong-An Hsu

National Central University, Chung-Li, Taiwan

Deng-Yuan Huang

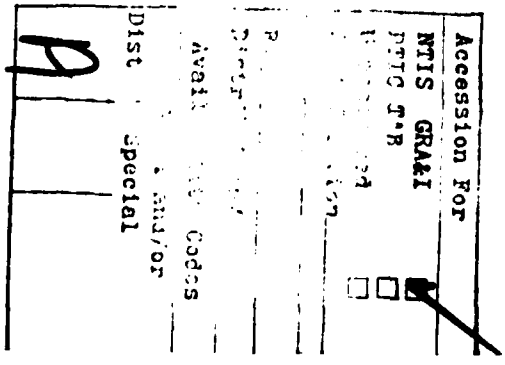
Academia Sinica, Taipei, Taiwan

ABSTRACT

In the past decade a number of fixed sampling methods have been developed for selecting the "best" or at least a "good" subset of variable in regression analysis. We are interested in deriving a sequential selection procedure to select a subset of a random size including all good regression equations. Tables for an example are given at the end of this paper.

1. INTRODUCTION

In the past decade a number of fixed sampling methods have been developed for selecting the "best" or at least a "good" subset of variables in regression analysis (see e.g. Arvesen and McCabe (1975) and Spjøtvøll (1972)). In this paper, we are interested in deriving a sequential selection procedure to select a random size subset including all "good" regression equations.



Tables for an example are given at the end of this paper.

2. SEQUENTIAL SUBSET SELECTION PROCEDURE

Before discussing the regression problem, we develop general results applicable to the selection of "good" or "superior" populations defined later.

Let $\pi_0, \pi_1, \ldots, \pi_k$ denote k+1 normal populations with unknown means $\mu_0, \mu_1, \ldots, \mu_k$ and variances $\sigma_0^2, \sigma_1^2, \ldots, \sigma_k^2$. Assume that σ_0^2 is known but σ_i^2 $(1 \leq i \leq k)$ are unknown. Let the ranked values of σ_i^2 be denoted by $\sigma_{[1]}^2 \leq \ldots \leq \sigma_{[k]}^2$. We wish to derive a method to construct a sequential procedure to select a subset containing all "superior" populations - the populations with smaller variances, with a probability not less than P*, $(0 < P^* < 1)$, a specified constant. We assume that $\sigma_0^2 = 1$.

Let X_{in} denote the nth observation from population π_i. It is assumed that the observations $X_{i1}, \ldots, X_{in}$ are independent random variables. Define

$$\bar{X}_{in} = \frac{1}{n} \sum_{j=1}^{n} X_{ij},$$

and

$$s_{in}^2 = \frac{1}{n-1} \sum_{j=1}^{n} (X_{ij} - \bar{X}_{in})^2.$$

The selection procedure will depend upon $\{s_{in}^2\}$ which is a sufficient and transitive sequence and also invariantly sufficient for $\{\sigma_i^2\}$.

A population π_i is said to be "superior" (or "good") if $\sigma_i^2 \leq \Delta$, to be "inferior" (or "bad") if $\sigma_i^2 > \Delta$, where Δ is a specified constant greater than 1. Let Ω be the parameter space which is the collection of all possible parameter vectors $\underline{\theta} = (\sigma_1^2, \ldots, \sigma_k^2)$. Let t denote the unknown number of inferior

populations in the given collection of k populations. We have $0 \leq t \leq k$. Let

$$\Omega_t = \{\underline{\theta} | \sigma^2_{[1]} \leq \cdots \leq \sigma^2_{[k-t]} < \Delta < \sigma^2_{[k-t+1]} \leq \cdots \leq \sigma^2_{[k]}\}.$$

Then $\Omega = \bigcup_{t=0}^{k} \Omega_t$.

For the subset selection procedure R, two constants Δ and P* with $\Delta > 1$, $1 > P^* > 0$, are specified and we wish to select a subset containing all superior populations with a probability of at least P*. When all the superior populations are contained in the selected subset, we say a correct decision (CD) has been made. Thus we require a procedure for which

$$P_{\underline{\theta}}(CD|R) \geq P^*$$

for all $\underline{\theta} \in \Omega$.

Let $g_{\sigma_i^2}(s^2_{in})$ denote the probability density of s^2_{in} depending on the parameter σ_i^2. We define the log-likelihood ratios

$$\ell_n(s^2_{in}) = \log \; g_{\Delta}(s^2_{in}) - \log \; g_1(s^2_{in}) \qquad (2.1)$$

upon which the procedure is based.

Elimination type sequential selection procedure R for selecting the superior populations.

Begin by taking n_1 (≥ 1) independent observations from each of the k populations. Calculate the values of the k log-likelihood ratios $\ell_{n_1}(s^2_{in_1})$, $1 \leq i \leq k$. For any i, if

$$\ell_{n_1}(s^2_{in_1}) \geq a,$$

where $a = \log(k(k+1)/2(1-P^*))$, we eliminate the population π_i from further consideration. We proceed to the next (second) stage by taking $n_2 - n_1$ independent observations on each of the remaining populations. The log-likelihood ratios for the contending populations are again computed and the same elimination rule is used

except that $\ell_{n_2}(s^2_{in_2})$ everywhere replaces $\ell_{n_1}(s^2_{in_1})$. We continue in this manner until the elimination is stopped, at time the procedure is terminated with the declaration that the remaining populations are the superior populations. If after applying this rule at the sth stage (say), the number of remaining populations is zero, then we select the population π_0 which is the control population.

Note that n_1 is the sample size of that stage of the procedure at which a decision may be made, for the first time, to reject one or more populations. Let $n_2 > n_1$ be the sample size of the next stage of the procedure at which such a decision may be made, and in general let $n_s > n_{s-1}$ be the sample size of the stage of the procedure at which the sth decision to reject one or more populations may be made. Let N be the stage at which the procedure terminates. It is clear that if there are k populations to start with, then $N \leq n_k$ (see Gupta and Huang (1975)).

We assume that

$$P_{\sigma_i^2}\{\ell_n(s^2_{in}) \geq a \quad \text{for some } n\} \tag{2.2}$$

is a nondecreasing function of σ_i^2. A sufficient condition for this is discussed by Hoel (1970). Without loss of generality, we assume that $\pi_1,\ldots,\pi_{k-t}$ are the superior populations. Since the procedure R is truncated, we have

$$\begin{aligned}
1-P(CD|R) &= P\{\ell_n(s^2_{in}) \geq a \quad \text{for some } i = 1,\ldots,k-t, \\
&\quad \text{for some} \quad t,\ 0 < t \leq k, \quad \text{for some } n\} \\
&= \sum_{t=0}^{k} \sum_{i=1}^{k-t} P_{\sigma_i^2}\{\ell_n(s^2_{in}) \geq a \quad \text{for some } n\} \\
&< \sum_{t=0}^{k} \sum_{i=1}^{k-t} P_{\Delta}\{\ell_n(s^2_{in}) \geq a \quad \text{for some } n\} \\
&= \sum_{t=0}^{k} (k-t)e^{-a} - \frac{1}{2}k(k+1)e^{-a} = 1-P^*.
\end{aligned}$$

3. APPLICATIONS TO SELECTION OF "GOOD" OR "SUPERIOR" REGRESSION EQUATIONS

Assume the following standard linear model as follows,

$$Y = X\beta + \varepsilon \tag{3.1}$$

where X is an nxp known matrix of rank $p \leq n$, β is a pxl parameter vector, and $\varepsilon \sim N(0, \sigma_0^2 I_n)$. Consider the models for any r, $2 \leq r \leq p-1$,

$$Y = X_{ri}\beta_{ri} + \varepsilon_{ri} \tag{3.2}$$

where X_{ri} is an nxr matrix of rank r with $X_{11}' = [1,\ldots,1]_{1Xn}$, β_{ri} is a rxl parameter vector, and $\varepsilon_{ri} \sim N(0, \sigma_{ri}^2 I_n)$, where $i=1,\ldots,k_r = \binom{p-1}{r-1}$. Let $k=\sum_{r=2}^{p-1} k_r$. The goal is to include all the designs X_{ri} (or sets of independent variables) associated with $\sigma_{[j]}^2$, $j = 1,\ldots,k-t$.

Note that for any r, $2 \leq r \leq p-1$, if

$$SS_{ri} = Y'\{I - X_{ri}(X_{ri}'X_{ri})^{-1}X_{ri}'\}Y = Y'Q_{ri}Y,$$

then following Searle (1972, p. 57)

$$SS_{ri}/\sigma_0^2 \sim \chi^2\{\nu_r, (X\beta)'Q_{ri}(X\beta)/(2\sigma_0^2)\},$$

where $\nu_r = n-r$, for $1 \leq i \leq k_r$. Note that the noncentrality parameter is not zero in general and that

$$\sigma_{ri}^2 = \sigma_0^2 + (X\beta)'Q_{ri}(X\beta)/\nu_r.$$

If σ_0^2 is not equal to 1, then we consider the linear model $Y/\sigma_0 = X\beta/\sigma_0 + \varepsilon'$, $\varepsilon' \sim N(0, I_n)$. Thus we assume without loss of generality that $\sigma_0^2 = 1$.

We know that the non-central $\chi^2(x,\lambda)$ with non-centrality parameter λ has monotone likelihood ratio in x. Hence the monotonicity of (2.2) is satisfied. We can apply the sequential procedure R to select superior regression equations by replacing $s_{ri,n}^2$ by SS_{ri}/ν_r.

4. COMPUTATION OF (2.1)

Let $U_{ri} = SS_{ri}/\nu_r$. The probability density of U_{ri} is

$$g_{\sigma_{ri}^2}(U_{ri}) = \nu_r \left[e^{-\lambda} \sum_{k=0}^{\infty} \frac{\lambda^k (\nu_r U_{ri})^{\frac{1}{2}\nu_r + k - 1} e^{-\frac{1}{2}(\nu_r U_{ri})}}{k!\, 2^{\frac{1}{2}\nu_r + k}\, \Gamma(\frac{1}{2}\nu_r + k)} \right]$$

where $\sigma_{ri}^2 = 1 + (X\beta)'Q_{ri}(X\beta)/\nu_r$, $\nu_r = n-r$ and $\lambda = (X\beta)'Q_{ri}(X\beta)/2$. If $\sigma_{ri}^2 = 1$, then $\lambda = 0$ and if $\sigma_{ri}^2 = \Delta$, then $\lambda = (\Delta-1)\nu_r/2$. Hence

$$q_2(U_{ri})/q_1(U_{ri}) = e^{-\lambda} \sum_{k=0}^{\infty} \frac{\lambda^k}{k!} \left[\frac{\nu_r U_{ri}}{2} \right]^k \frac{\Gamma(\frac{1}{2}\nu_r)}{\Gamma(\frac{1}{2}\nu_r + k)} \qquad (4.1)$$

where $\lambda = (\Delta-1)\nu_r/2$. Let

$$a_k = \frac{e^{-\lambda}\lambda^k}{k!} \left[\frac{\nu_r U_{ri}}{2} \right]^k \frac{\Gamma(\frac{1}{2}\nu_r)}{\Gamma(\frac{1}{2}\nu_r + k)}, \quad k = 0,1,2,\ldots$$

Since

$$\frac{a_{k+1}}{a_k} = \frac{\lambda}{k+1} \left[\frac{\nu_r U_{ri}}{2} \right] \frac{1}{(\frac{1}{2}\nu_r + k)} \to 0 \quad \text{as } k \to \infty,$$

then for any $0 < \delta < 1$, there exists q such that

$$\frac{a_{k+1}}{a_k} \leq \delta < 1, \quad \text{for all } k \geq q.$$

Let us consider the error due to the truncation of the series in (4.1). Let q be the number of terms in the truncated series. Then the error due to truncation of the series in (4.1) is given by

$$\sum_{k=0}^{\infty} a_{q+k} \leq \frac{a_q}{1-\delta}.$$

Given $\eta > 0$, let k_0 be the smallest positive integer k such that

$$\frac{a_k}{\eta} < 1 \quad \text{and} \quad \frac{a_{k+1}}{a_k} + \frac{a_k}{\eta} \leq 1.$$

For this k_0, it is easy to prove that

$$0 < g_\Lambda(U_{ri})/g_1(U_{ri}) - \left[\sum_{k=0}^{k_0-1} a_k\right] = \sum_{k=0}^{\infty} a_{k_0+k} \leq \eta.$$

Thus $g_\Lambda(U_{ri})/g_1(U_{ri}) \doteq \sum_{k=0}^{k_0-1} a_k$ with error less than η. To evaluate $g_\Lambda(U_{ri})/g_1(U_{ri})$, the computation is very efficient.

5. EXAMPLE

In this section we present an example which will serve to illustrate the sequential subset selection procedure. The data set is taken from Neter and Wasserman (1974, p. 373), who used it to illustrate several methods of finding a "best" set of independent variables.

There are n = 55 observations on p = 5 independent variables. Then $k = 2^4 - 2 = 14$. For the subset selection procedure R, two constants Λ and P* with $\Lambda > 1$, $1 > P^* > 0$, are specified and we wish to select a subset containing all superior regression equations with probability at least P*.

Begin by taking n_1 (≥ 5) independent observations. Calculate the values of the k ratios $g_\Lambda(U_{ri})/g_1(U_{ri})$ with error η (specified). For any r, i, If

$$g_\Lambda(U_{ri})/g_1(U_{ri}) \geq b$$

where $b = k(k+1)/2(1-P^*)$, we eliminate the regression equation from further consideration. We proceed to the second stage by taking $n_2 - n_1$ independent observations on each of the remaining regression equations. The ratios are again computed and the same

TABLE I

Data for the Forest Inventory

Y	X_1	X_2	X_3	X_4	X_5	Y	X_1	X_2	X_3	X_4	X_5
200	1	170	47	50	3.617	82	1	67	21	70	3.19
358	1	310	30	70	10.333	112	1	75	24	50	3.125
167	1	131	26	50	5.038	24	1	11	7	50	1.571
111	1	83	19	50	4.368	185	1	165	31	80	5.323
290	1	231	38	70	6.079	301	1	279	26	80	10.731
36	1	21	9	50	2.333	284	1	306	26	90	11.769
133	1	80	27	50	2.963	467	1	477	30	90	15.9
135	1	91	23	50	3.957	410	1	368	23	90	16
399	1	369	67	90	5.507	279	1	224	34	30	6.588
279	1	256	25	90	10.24	182	1	151	29	50	5.207
280	1	321	42	90	7.643	126	1	109	25	50	4.36
146	1	104	33	30	3.512	160	1	148	41	50	3.61
123	1	88	34	30	2.588	141	1	112	18	70	6.222
122	1	86	30	30	2.867	31	1	21	13	30	1.615
103	1	84	29	30	2.897	78	1	56	33	30	1.697
264	1	239	52	70	4.596	45	1	29	18	40	1.611
360	1	323	51	90	6.333	93	1	72	37	50	1.946
55	1	45	18	30	2.5	220	1	201	47	60	4.277
421	1	389	50	70	7.78	207	1	170	33	50	5.152
89	1	61	31	50	1.968	170	1	140	27	50	5.185
98	1	58	6	30	9.667	92	1	69	33	40	2.091
262	1	239	45	90	5.311	99	1	79	9	50	8.778
147	1	126	17	80	7.412	85	1	67	20	40	3.35
287	1	273	29	70	9.414	136	1	125	15	80	8.333
73	1	83	17	80	4.882	653	1	583	28	60	20.821
86	1	63	23	50	2.739	111	1	89	26	70	3.423
104	1	60	31	50	1.935	22	1	18	4	40	4.5
330	1	310	65	50	4.769						

elimination rule is used. We continue in this manner until the elimination is stopped, at which time the procedure is terminated with the declaration that the remaining regression equations are the superior regression equations.

Let $\eta = 0.1$. For the value of $g_\Delta(U_{ri})/g_1(U_{ri})$, this error of $\eta = 0.1$ is small enough with respect to constant b. Table II-VII are the subsets of independent variables of elimination for the sequential subset selection procedure R.

Table III, we consider $\Delta = 1.2$. If $P^* = 0.9$, then the procedure R eliminates (X_1, X_5), (X_1, X_4) and (X_1, X_3) at stage 1 $(n_1 = 11)$; eliminate (X_1, X_4, X_5) at stage 2 $(n_2 = 16)$ and eliminate (X_1, X_3, X_4) at stage 3 $(n_3 = 21)$. No subset is eliminated at stage 4 $(n_4 = 26)$. Thus the procedure is terminated. (X_1, X_2), (X_1, X_2, X_3), (X_1, X_2, X_4), (X_1, X_2, X_5), (X_1, X_3, X_5), (X_1, X_2, X_3, X_4), (X_1, X_2, X_3, X_5), (X_1, X_2, X_4, X_5) and (X_1, X_3, X_4, X_5) are the set of variables of superior regression equations. We can use C_p statistic to select one of good regression equations among the set of superior regression equations. For this example, (X_1, X_2, X_4) is the set of variable of a good regression equation (cf. Neter and Wasserman (1974)). Table II-VII represents the results for $\Delta = 1.1$, 1.2, 1.5, 2, 3 and 5; $P^* = 0.7$, 0.8 and 0.9.

TABLE II

$\eta = 0.1$, $\Delta = 1.1$.

P* \ n	16	21	26	31
0.7	(1.4),(1,3) (1,5)	(1,4,5),(1,3.4)	no rejection	———
0.8	(1,5),(1,3)	(1,4.5),(1,3,4) (1,4)	no rejection	———
0.9	(1,5),(1,3)	(1,4,5),(1.4)	(1,3,4)	no rejection

TABLE III

$\eta = 0.1$, $\Delta = 1.2$.

p* \ n	11	16	21	26
0.7	(1,4,5),(1,5) (1,4),(1,3)	no rejection	———	———
0.8	(1,4,5),(1,5) (1,4),(1,3)	no rejection	———	———
0.9	(1,5),(1,4) (1,3)	(1,4,5)	(1,3,4)	no rejection

TABLE IV

$\eta = 0.1$, $\Delta = 1.5$.

p* \ n	6	11	16
0.7	(1,4),(1,3)	(1,4,5),(1,5) (1,3,4)	no rejection
0.8	(1,3)	(1,4,5),(1,5) (1,3,4),(1,4)	no rejection
0.9	(1,3)	(1,4,5),(1,5) (1,3,4),(1,4)	no rejection

TABLE V

$\eta = 0.1$, $\Delta = 2$.

p* \ n	6	11	16
0.7	(1,4,5),(1,5) (1,4),(1,3)	(1,3,4)	no rejection
0.8	(1,5),(1,4) (1,3)	(1,4,5),(1,3,4)	no rejection
0.9	(1,5),(1,4) (1,3)	(1,4,5),(1,3,4)	no rejection

TABLE VI

$\eta = 0.1$, $\Delta = 3$.

p* \ n	6	11	16
0.7	(1,4,5),(1,5),(1,3,4) (1,4),(1,3)	no rejection	———
0.8	(1,4,5),(1,5) (1,4),(1,3)	(1,3,4)	no rejection
0.9	(1,4,5),(1,5) (1,4),(1,3)	(1,3,4)	no rejection

TABLE VII

$\eta = 0.1$, $\Delta = 5$.

p* \ n	6	11
0.7	(1,4,5),(1,5),(1,3,4) (1,4),(1,3)	no rejection
0.8	(1,4,5),(1,5),(1,3,4) (1,4),(1,3)	no rejection
0.9	(1,4,5),(1,5),(1,3,4) (1,4),(1,3)	no rejection

ACKNOWLEDGEMENT

This research was supported by the Office of Naval Research contract N00014-75-C-0455 at Purdue University. Reproduction in whole or in part is permitted for any purpose of the United States Government.

The authors would also like to thank the National Science Council of Republic of China for some financial support.

BIBLIOGRAPHY

Arvesen, J. N. and McCabe, G. P. Jr. (1975). Subset selection problems for variances with applications to regression analysis. *J. Amer. Statist. Assoc.* 70, 166-170.

Gupta, S. S. and Huang, D. Y. (1975). On some parametric and nonparametric sequential subset selection procedures. *Statistical Inference and Related Topics*, M. L. Puri, Ed., Vol. 2, 101-128.

Hoel, D. G. (1970). On the monotonicity of the OC of an SPRT. *Ann. Math. Statist.* 41, 310-314.

Neter, J. and Wasserman, W. (1974). *Applied Linear Statistical Models*, Vol. 1, Richard D. Irwin, Inc.

Searle, S. R. (1971). *Linear Models*. New York, John Wiley and Sons, Inc.

Spjøtvøll, E. (1972). Multiple comparison of regression functions. *Ann. Math. Statist.* 43, 1076-1088.

Key Words and Phrases: Sequential, Selection procedures, Regression analysis, Superior, Inferior populations.

SECURITY CLASSIFICATION OF THIS PAGE *(When Data Entered)*

REPORT DOCUMENTATION PAGE		READ INSTRUCTIONS BEFORE COMPLETING FORM
1. REPORT NUMBER Mimeograph Series #81-21	2. GOVT ACCESSION NO.	3. RECIPIENT'S CATALOG NUMBER
4. TITLE *(and Subtitle)* Some Sequential Selection Procedures for Good Regression Models		5. TYPE OF REPORT & PERIOD COVERED Technical 6. PERFORMING ORG. REPORT NUMBER Mimeo. Series #81-21
7. AUTHOR(s) Tong-An Hsu and Deng-Yuan Huang		8. CONTRACT OR GRANT NUMBER(s) ONR N00014-75-C-0455
9. PERFORMING ORGANIZATION NAME AND ADDRESS Purdue University Department of Statistics West Lafayette, Indiana 47907		10. PROGRAM ELEMENT, PROJECT, TASK AREA & WORK UNIT NUMBERS
11. CONTROLLING OFFICE NAME AND ADDRESS Office of Naval Research Washington, DC		12. REPORT DATE June 1981 13. NUMBER OF PAGES 12
14. MONITORING AGENCY NAME & ADDRESS*(if different from Controlling Office)*		15. SECURITY CLASS. *(of this report)* Unclassified 15a. DECLASSIFICATION DOWNGRADING SCHEDULE

16. DISTRIBUTION STATEMENT *(of this Report)*

Approved for public release, distribution unlimited.

17. DISTRIBUTION ST[illegible]ENT *(of the abstract entered in Block 20, if different from Report)*

18. SUPPLEMENTARY [illegible]TES

19. KEY WORDS *(Continue on reverse side if necessary and identify by block number)*

Sequential, Selection procedures, Regression analysis, Superior, Inferior populations.

20. ABSTRACT *(Continue on reverse side if necessary and identify by block number)*

In the past decade a number of fixed sampling methods have been developed for selecting the "best" or at least a "good" subset of variable in regression analysis. We are interested in deriving a sequential selection procedure to select a subset of a random size including all good regression equations. Tables for an example are given at the end of this paper.

DD FORM 1 JAN 73 1473

UNCLASSIFIED


DAT
FILM